Cross Language Information Retrieval (CLIR) Techniques using Query Optimization on Marathi Language

Synopsis Submitted to

Swami Ramanand Teerth Marathwada University
for registration to Ph.D. degree (Computer Science)
in the faculty of Science

by

Suhas Dnyandeo Pachpande
Department of Computer Science
Sant Gadge Baba Amravati University
Amravati

Supervisor

Dr. Parag Bhalchandra
Assistant Professor
School of Computational Sciences
Swami Ramanand Teerth Marathwada University

Swami Ramanand Teerth Marathwada University, Nanded

July 2019

Title:

Cross Language Information Retrieval (CLIR) Techniques using Query Optimization on Marathi Language

Aim:

The aim of research is to propose a model for improving efficiency and relevance of information retrieval across the language barriers using methods for Cross Language Information Retrieval (CLIR).

Objectives:

Objectives of the proposed research are as follows:

1. To analyze the scope of information retrieval across various language domains.
2. To investigate and analyze the existing approaches for Cross Language Information Retrieval.
3. To critically study various translation approaches and examine their effectiveness in cross language domains.
4. To understand various models like Boolean, Vector Space and Probabilistic models of information retrieval.
5. To identify standard data sets in Marathi language, if available.
6. To create a dataset for Marathi language, if standard dataset is not available.
7. To apply the models mentioned above for cross language information retrieval.
8. To explore and design a model like VSM which is more relevant to enhance the results of information retrieval.

Introduction

Searching information has become an integral part of our life. Generally we search in a particular language (usually in English). The information we are interested in may also be available in other languages. This has cultivated the concept of Cross-Language Information Retrieval (CLIR) that allows retrieval of documents across language boundaries. i.e. It deals with the problem of requesting the information in one language while retrieving results and documents in one or several other languages [1].

The search models usually prepare an internal representation for document retrieval based on the query which serves as a core component of information retrieval process. For retrieving documents from multiple languages, a single internal representation is naturally not sufficient. A data fusion approach offering two or more document representations may be able to produce better results [2]. Besides the problems of monolingual information retrieval, translation is the key problem in CLIR because of the multiple representations of text.

Real Challenge is to retrieve same or similar internal representation for documents containing similar piece of information, but written in different languages.

In monolingual IR, a standard representation space is created by word stemming forming terms for comparisons. The primary objective of CLIR is to build tools that would help retrieving equivalent and associated terms from other languages having same meaning.

Current Work :

Primarily CLIR is performed by query translation which is further followed by monolingual retrieval. Thus ultimately the retrieval performance is depends on the quality and effectiveness of query translation. Research communities working in this area are primarily focusing on following approaches:

Dictionary-based translation: Uses bilingual dictionary and selects the possible translations of each source word. These translated words collectively form a representation of the user's query in the target language. Compounding is commonly used process in almost all languages to form new words by combining two or more words. It keeps adding several new words to the language making it difficult to be accommodated and included any dictionary for that language [3]. Thus even though the retrieval results using dictionary translation approach are pretty good, they are still affected due to ambiguous translations leading more scope towards further research.

Machine translation: This strategy suggests translation of entire documents and web pages before indexing. An MT system can be used to translate the query, the document, or both into the same language, and retrieval process could then be treated with a general IR system [4].

This would reduce time consumed for query translation and may effectively retrieve documents in cross languages. But practically it is very difficult and computationally expensive to translate all documents in many probable languages.

Translation based on parallel corpora: A corpus is designed for both the source language texts and the target language. This approach attempts to extract strong translation relations between the two languages. This relation can be either at word level or at a higher level like phrases, classification and categorization. These relations are used to translate queries and documents. Collecting parallel texts of different language versions from various sources like the Web and building a corpus may make mining of parallel texts feasible and easier [5]. But designing corpus for each language is still limited to a smaller domain leading scope of further improvements.

Some of the popular search engines (like google and yahoo) partially provide multilingual search where the users query is translated to target language and it attempts to translate the retrieved documents back to source language using Machine Translation (MT) system. But CLIR is in a state of infancy for these search engines leading scope of further improvements. The mean average precision (MAP) of CLIR was much lower than monolingual IR initially and is improving over the new techniques.

Considering the methods and techniques proposed so far for cross language information retrieval, it can be inferred that there is a scope for enhancement in these strategies yielding better results. To examine and evaluate these strategies across language domains, some language is to be selected. Considering the worldwide scope and historical aspects of Marathi language and its ever-growing volume in literature, Marathi may be one of the best alternative for choosing as dataset. Marathi language is rich in terms of alphabets, consonants and vowels and many of its words have multiple meanings that may lead to ambiguity. The words may have different meanings based on the context in which these words are used. These factors may impose several challenges during research and moreover very less work has been performed in this direction for cross language information retrieval. This is the major motive towards selection of the topic for this research.

Literature Survey

Jian-Yun Nie[6] has described describe an approach to CLIR based on parallel texts using statistical translation models. He has experimented with translation models trained on a manually prepared parallel corpus for mining parallel Web pages and their utilization for CLIR. He has found that when a query is translated word by word, the accuracy maybe low. He has concluded that if compound terms are extracted and incorporated into the translation models and translated as a unit, this can increase the translation accuracy.

Robert Krajewski, et al [7] have proposed a method for translation of a dictionary by means of mining repositories in the source and target languages. This method consists of two stages: (1) translation by lexical similarity, where words are compared graphically, and (2) translation by semantic similarity, where contexts are compared. The method and its phases are thoroughly analyzed. It was found that the size of the seed dictionary influences the quality of the phase of semantic translation.

As many researchers have pointed out that for dictionary-based cross-language information retrieval, the lack of comprehensive dictionaries for multiple different languages is the real bottleneck. Korn´el Mark´o, et al [8] have introduced a methodology by which multilingual dictionaries are constructed automatically from simple seed lexicons by cognate mapping, from previously manually constructed sources. This method validates the lexical and semantic hypotheses and iteratively generates the new ones with the help of co-occurrence patterns of hypothesized translation synonyms in parallel corpora. The factor that limits the generality of this approach is that it relies on large aligned thesauri which may not be available all the time for all the languages.

Djoerd Hiemstra [9], et al have evaluated the tools and methods for Cross Language Information Retrieval (CLIR) with the TREC CLIR task document collection using Dutch queries on the English document base. They have addressed the two approaches to disambiguation, finding the correct translation for each query term before searching and searching with more than one possible translation. Through their experimental study, they have suggested that quality of search methods is more important than the quality of disambiguation methods and have found that good retrieval methods are able to disambiguate translated queries implicitly during searching.

Wessel Kraaij, et al [10] have investigated the problem of automatically mining parallel texts from the Web and different ways of integrating the translation models within the retrieval process. Through experiments, they have shown that the Web-based translation models can surpass commercial MT systems in CLIR tasks opening the perspective of constructing a fully automatic query translation device for CLIR at a very low cost. Although this approach is fully automated which avoid the tedious work of manual collection of parallel corpora, it was found that the mining results contain more noise (i.e., nonparallel pages).

Dan Wu, et al [11] has introduced the concept of Interactive Cross-language Search Engine with Translation Enhancement. It is stated that Relevance feedback (RF) obtained from users is found to be effective in improving retrieval results. The authors have suggested a novel RF approach called translation enhancement (TE) and integrated with the Query Expansion (QE) approach in CLIR. The implementation of approach is limited to CLIR for English queries over Chinese documents collection. It is further stated that with a machine translation system between document and query language may improve the results.

Mandeep Pannu, et al [12] have presented a comparison of various models for information retrieval like Boolean Information Retrieval (BIR), Vector Space Model (VSM) and Probabilistic Information Retrieval (PIR). Vector Space Model (VSM) is an algebraic model that offers a way of representing and comparing documents and queries based on keywords and provide a unique feature of ranking similarity between documents. Although Boolean Retrieval model is simple and convenient, it is more restrictive in the formulation of similarities i.e. in two states only: true or false. The probabilistic approach models the subjective features of information retrieval process over a range of probabilities but calculation of probabilities requires specification of assumptions that can be highly biased and inconsistent. The Vector Space Model combines clarity with flexibility and is well specified and well understood as the documents are represented with more details. Thus the VSM offers a viable compromise in information retrieval processing.

Krister Lind´en [13] has described the issue associated with new technical terms having a common origin rendered with different spelling of the underlying sounds, also known as cross-lingual spelling variants (CLSV). He has presented the problem of finding cross-lingual spelling variants in a probabilistic framework.

Lei Zhang, et al [14] have analyzed the importance of semantic relatedness in the process of cross language information retrieval for processing natural language texts composed in different languages. Cross-lingual semantic relatedness measures the strength of semantic connection between documents (or other textual units such as words, sentences and paragraphs) in different languages. The authors have described various approaches to semantic relatedness like dictionary-based approaches, thesaurus-based approaches and corpus-based approaches.

Proposed Work/Methodology:

Various IR models such as Boolean, Vector-space, Probabilistic models have been identified for the proposed research work.

For the language MARATHI, standard datasets, if available, will be identified. However, most of the researchers working in the domain of CLIR have created their own domain specific data set due to the non-availability of standard data sets. These data sets are not sufficient to be treated as standard sets for generalized study. Hence creation of Marathi data set will be one of the inevitable aspect of the proposed research.

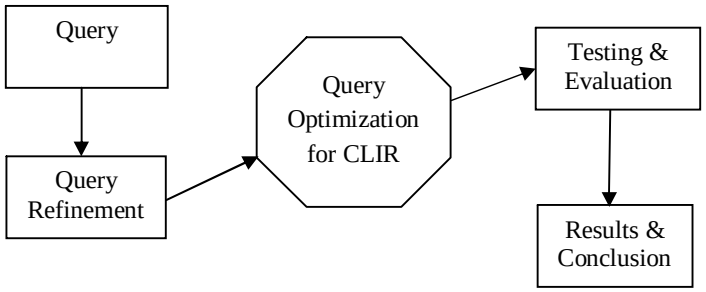
Researchers have shown by experimentation that Fuzzy Matching has been found to be effective in CLIR wherein concept based retrieval can be done using fuzzy matching. But this method has its own limitations.

As can be inferred from the study conducted through literature survey, the VSM seems to be most resulting over other models after testing it on standard dataset for monolingual IR. However, other methods will also be evaluated empirically. The VSM method will be explored with experimentation on standard data sets. Results with both traditional IR methods with VSM method will be compared.

The probable stages in proposed research work would be as follows:

- Literature Review
- Problem Formulation
- Identification of standard dataset / Creation of new dataset
- Examining the effectiveness of translation approaches in CLIR
- Study of various models for information retrieval
- Designing a model for CLIR
- Application of these models on identified dataset
- Testing and evaluation of proposed model on identified dataset
- Compilation of results and deriving conclusion
- Work of thesis writing

Probable Model for CLIR:



References:

[1] Monika Sharma, Sudha Morwal, "A Survey on Cross Language Information Retrieval", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015, Pages 384-387

[2] Jacques Savoy, "Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval", Information Retrieval, Vol. 7 (2004), Pages 121-148

[3] Turid Hedlund, "Dictionary-based Cross-Language Information Retrieval", Thesis submitted for Academic dissertation in the University of Tampere (Nov 2003), Pages 1-84

[4] Peishan Tsai, "An Introduction to Cross-Language Information Retrieval Approaches", LIS 531 - Information Retrieval, Pages 1-20

[5] Pu-Jen Cheng, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu, Lee-Feng Chien, "Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval", SIGIR'04 (ACM), July 25–29, 2004, Pages 146-153

[6] Jian-Yun Nie, "Cross-Language Information Retrieval", IEEE Computational Intelligence Bulletin, Vol.2 No.1 June 2003, Pages 19-23,

[7] Robert Krajewski, Henryk Rybinski, Marek Kozlowski, "A novel method for dictionary translation", Journal of Intelligent Information Systems (Springer) 2016, 47, Pages 491-514

[8] Korn´el Mark´o, Stefan Schulz, Olena Medelyan, Olena Medelyan, "Bootstrapping Dictionaries for Cross Language Information Retrieval", SIGIR'05 (ACM), August 15–19, 2005, Pages 528-535

[9] Djoerd Hiemstra, Franciska de Jong, "Disambiguation Strategies for Cross-Language Information Retrieval", ECDL'99, LNCS 1696 (Springer 1999), Pages 274-293

[10] Wessel Kraaij, Jian-Yun Nie, Michel Simard, "Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval", Computational Linguistics Volume 29, Number 3, Pages 381-419

[11] Dan Wu, Daqing He, "ICE-TEA: an Interactive Cross-language Search Engine with Translation Enhancement", SIGIR'08 (ACM), July 20–24, 2008, Page 882

[12] Mandeep Pannu, Anne James, Robert Bird, "A Comparison of Information Retrieval Models, WCCCE'14 (ACM), May 2-3, 2014

[13] Krister Lind´en, "Multilingual modeling of cross-lingual spelling variants", Journal of Information Retrieval (Springer) 2006, Pages 295-310

[14] Lei Zhang, Thanh Tran, Achim Rettinger, "A Theoretical Analysis of Cross-lingual Semantic Relatedness in Vector Space Models, ICTIR'15 (ACM), September 27–30, Pages 241-150

Research Supervisor

Dr. Parag Bhalchandra
Assistant Professor
School of Computational Sciences
S.R.T.M. University, Nanded

Research Student

Suhas D. Pachpande